

MS2a, Model Solutions Week 4

Rune Lyngsø

November 5, 2009

A Phylogeny Reconstruction

- a. Consider a binary character – for convenience denote the two possible states 0 and 1 – evolving according to the rate matrix

$$Q = \begin{bmatrix} -\alpha & \alpha \\ \alpha & -\alpha \end{bmatrix}$$

Determine $P(t) = e^{Qt}$.

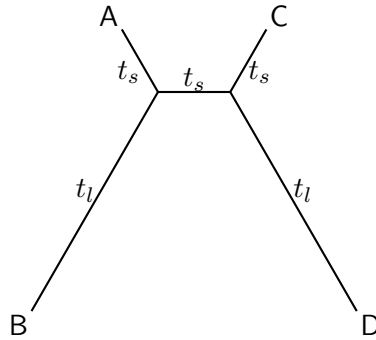
This is very similar to the problem on the problem sheet for week 2, except that we only have two possible characters. We can proceed in the same manner as then, or we could observe that the characteristic polynomial is $(-\alpha - \lambda)^2 - \alpha^2 = \lambda^2 + 2\alpha\lambda$. Hence Q has eigenvalues 0 and -2α . This immediately tells us that Q has two independent eigenvectors and thus can be diagonalised, which again means that

$$P(t) = e^{Qt} = B \begin{bmatrix} 1 & 0 \\ 0 & e^{-2\alpha t} \end{bmatrix} B^{-1}$$

for some properly chosen matrix B . Without even finding B we know that $P(t)_{00} = a + be^{-2\alpha t}$ for some $a, b \in \mathbf{R}$. We further know that $P(0)_{00} = 1 \Rightarrow a + b = 1$ and $P'(0)_{00} = -\alpha \Rightarrow b = 1/2$. We can conclude that $a = b = 1/2$. Identical reasoning can be used for $P(t)_{11}$, and $P(t)_{01} = 1 - P(t)_{00}$ and $P(t)_{10} = 1 - P(t)_{11}$ completes the picture to give

$$P(t) = \begin{bmatrix} \frac{1}{2} + \frac{1}{2}e^{-2\alpha t} & \frac{1}{2} - \frac{1}{2}e^{-2\alpha t} \\ \frac{1}{2} - \frac{1}{2}e^{-2\alpha t} & \frac{1}{2} + \frac{1}{2}e^{-2\alpha t} \end{bmatrix}$$

- b. Assume that we have a sequence of this binary character evolving on the following tree



with observed sequences A, B, C, and D. Let t_s be chosen such that $P(t)_{01} = 1/20$ and t_l be chosen such that $P(t)_{01} = 1/4$. What are the values of αt_s and αt_l meeting this requirement?

$$P(t_s)_{01} = \frac{1}{2} - \frac{1}{2}e^{-2\alpha t_s} = \frac{1}{20} \Rightarrow \alpha t_s = \ln \frac{\sqrt{10}}{3}$$

$$P(t_l)_{01} = \frac{1}{2} - \frac{1}{2}e^{-2\alpha t_l} = \frac{1}{4} \Rightarrow \alpha t_l = \ln \sqrt{2}$$

- c. What are the probabilities of observing each of the 16 possible combinations of the binary character at the four sequences, *i.e.* the probability of observing a 0 in all four sequences, a 0 in sequences A, B, and C and a 1 in sequence D, *etc.*?

Due to the symmetry of the evolutionary model, the patterns come in pairs with identical probability, *e.g.* $p_{0100} = p_{1011}$, so we only need to calculate probabilities for, say, patterns with 0 observed at A. With just two internal nodes it is as easy enumerating the four possible combinations of characters in these nodes as anything else. This gives

	00	00	00	00	p
0000	$\frac{61,731}{256,000}$	$\frac{57}{256,000}$	$\frac{57}{256,000}$	$\frac{19}{256,000}$	$\frac{7,733}{32,000}$
0001	$\frac{20,577}{256,000}$	$\frac{171}{256,000}$	$\frac{19}{256,000}$	$\frac{57}{256,000}$	$\frac{2,603}{32,000}$
0010	$\frac{3,249}{256,000}$	$\frac{1,083}{256,000}$	$\frac{3}{256,000}$	$\frac{361}{256,000}$	$\frac{587}{32,000}$
0011	$\frac{1,083}{256,000}$	$\frac{3,249}{256,000}$	$\frac{1}{256,000}$	$\frac{1,083}{256,000}$	$\frac{677}{32,000}$
0100	$\frac{20,577}{256,000}$	$\frac{19}{256,000}$	$\frac{171}{256,000}$	$\frac{57}{256,000}$	$\frac{2,603}{32,000}$
0101	$\frac{6,859}{256,000}$	$\frac{57}{256,000}$	$\frac{57}{256,000}$	$\frac{171}{256,000}$	$\frac{893}{32,000}$
0110	$\frac{1,083}{256,000}$	$\frac{361}{256,000}$	$\frac{9}{256,000}$	$\frac{1,083}{256,000}$	$\frac{317}{32,000}$
0111	$\frac{361}{256,000}$	$\frac{1,083}{256,000}$	$\frac{3}{256,000}$	$\frac{3,249}{256,000}$	$\frac{587}{32,000}$

- d. For sequence length $n \rightarrow \infty$ what tree would you expect to be preferred by the parsimony method, *i.e.* the tree requiring the fewest character changes when summed over all sequence positions?

The pattern 0011 supports the tree grouping sequences A and B, the pattern 0101 supports the tree grouping sequences A and C together, while the pattern 0110 supports the tree grouping sequences A and D together. The remaining patterns (with 0 observed at A) do not distinguish between the three possible topologies. The 0101 pattern will occur the most for sufficiently long sequences (about a third more occurrences than the pattern 0011 and almost three times as often as the pattern 0110) so we would expect the parsimony method to choose the tree grouping sequences A and C together.

- e. Write an expression in terms of the probabilities of the 16 possible character combinations (*i.e.* your variables will be $p_{0000}, \dots, p_{1111}$) that should be maximised to find the phylogeny the maximum likelihood method will converge to for $n \rightarrow \infty$?

The likelihood is just the probability of the data given the parameters (topology and rates, here represented by the probabilities of the 16 possible character combinations). For $n \rightarrow \infty$ we know that the fraction of columns observed with a particular character combination will converge to the probability of observing that combination, which we computed above. So the likelihood expression we should maximise is

$$\left(p_{0000}^{7733} p_{0001}^{2603} p_{0010}^{587} p_{0011}^{677} p_{0100}^{2603} p_{0101}^{893} p_{0110}^{317} p_{0111}^{587} p_{1000}^{587} p_{1001}^{317} p_{1010}^{893} p_{1011}^{2603} p_{1100}^{677} p_{1101}^{587} p_{1110}^{2603} p_{1111}^{7733} \right)^{n/32000}$$

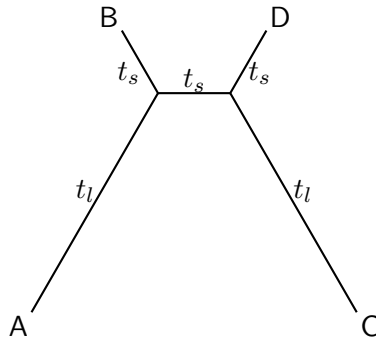
For the MLE we can ignore n as this will not change the location of the maximum (for non-negative likelihoods).

Without analytically solving for the MLE phylogeny, which phylogeny do you expect it to be?

We know that the maximum likelihood method converges to the true tree under this model, so the true phylogeny is the expected outcome. To further corroborate this claim, we can observe that the likelihood will be maximised for phylogenies where the probability of observing each character combination equals the observed fraction of this combination. This evidently holds for the phylogeny we have simulated

the data from. We can now write an equation system for this identity, and possibly using the assistance of *e.g.* Maple establish that the true phylogeny is the only solution to this equation system for probabilities of observed changes restricted to be between 0 and 1/2.

- f. Assume now that half the positions of our sequence evolve on the tree above, and half the positions evolve on the following tree



that is the tree where A and C sit at the end of long branches instead of B and D. What is now the probability of observation of the 16 possible patterns of character states at the four sequences?

Again it suffices to consider combinations with a 0 observed at sequence A. Apart from that, we just need to take the average of two probabilities from the model where only one tree was assumed, *e.g.* we get the probability of the pattern 0010 by taking the average of the probabilities of observing the patterns 0010 and 0001 under the one tree model. For some patterns, the two patterns averaged over are the same. The results are summarised by the following table:

0000	0001	0010	0011	0100	0101	0110	0111
0000	0001/0010	0010/0001	0011	0100/0111	0101	0110	0111/0100
$\frac{7,733}{32,000}$	$\frac{1,595}{32,000}$	$\frac{1,595}{32,000}$	$\frac{677}{32,000}$	$\frac{1,595}{32,000}$	$\frac{893}{32,000}$	$\frac{317}{32,000}$	$\frac{1,595}{32,000}$

What is the tree expected to be preferred by the parsimony method?

The probabilities of the three patterns distinguishing between the three possible topologies is unchanged from the previous case, so the parsimony method will still postulate the wrong topology of grouping sequence A with sequence C.

- g. If you were told that the correct topology is in fact not the topology the maximum likelihood method will converge to, which of the alternate topologies would be your guess for the one the maximum likelihood method converges to instead?

If we know the right topology is not the MLE, then the only two remaining possibilities are the one that group sequence A with sequence C and the one that group sequence A with sequence D. Sequence A and Sequence D are not evolutionary close in either tree, so the best guess would be that the MLE groups sequence A and sequence C. This is indeed what happens. The equations are beyond Maple's capabilities, but resorting to PAML, a software package for maximising probabilities of sequences under various substitution models we get log-likelihoods of $-161,388$ for grouping sequence A with sequence B, $-160,010$ for grouping sequence A with sequence C, and $-162,073$ for grouping sequence A with sequence D.

B Alignment of Sequences

- a. Consider the three sequences

```
AGTCGGACAATGTC
GGCGAAAATGTACTTC
GGCACAAGTCGTTCC
```

What is the longest sequence you can find that is a subsequence of *all* three sequences? For example, TTT is a subsequence of all three sequences, but CTTT is not (it is not a subsequence of the first sequence).

The sequence GGCAATGTC is one longest common subsequence of the three sequences.

- b. What is the shortest sequence you can find such that all three sequences are a subsequence of it?

The sequence AGGTCGAGACAATGTACGTTCC is one shortest common supersequence of the three sequences.

- c. Assume you are allowed three operations: changing one character (e.g. AGT \rightarrow ACT), deleting one character (e.g. ACT \rightarrow AC), and inserting one character (e.g. AC \rightarrow ATC). What is the best sequence you can find in terms of minimising the maximum number of changes required to obtain the three sequences above (also known as a median of the three sequences)?

The two sequences GGCGACAATGTCC and GGCGACAATGTTC both have distance at most 4 to any of the three sequences.

- d. An alignment of a set of sequences is a matrix with one row for each sequence, where each entry contains either a character or a gap (usually depicted by $-$). When ignoring the gaps in a row, the remaining entries yields the corresponding sequence. Each column is required to contain at least one (non-gap) character. So an alignment of the above sequences could start as

```

A G T C G ...
G G - C G ...
G G - C - ...

```

How many different alignments are there of three sequences with three characters each?

Each column in an alignment uses zero or one characters from each sequence, and at least one sequence has to contribute a character. With sequences of no characters, there is just one possible alignment, and we can't have alignments of sequences with a negative number of characters. These considerations lead to the following recursion for the number of alignments on three sequences with a , b , and c characters:

$$A(a, b, c) = \begin{cases} 0 & \text{if } \min\{a, b, c\} < 0 \\ 1 & \text{if } a = b = c = 0 \\ -A(a, b, c) + \sum_{i=0}^1 \sum_{j=0}^1 \sum_{k=0}^1 A(a-i, b-j, c-k) & \text{otherwise} \end{cases}$$

Solving this recursion for $a = b = c = 3$ we get 16,081 different alignments.